M^{3}TCM: Multi-modal Multi-task Context Model for Utterance Classification in Motivational Interviews

Sayed Muddashir Hossain, Jan Alexandersson, Philipp Müller
German Research Center for Artificial Intelligence
Saarbrücken, Germany
sayed_muddashir.hossain@dfki.de, jan.alexandersson@dfki.de, philipp.mueller@dfki.de

Abstract

Accurate utterance classification in motivational interviews is crucial to automatically understand the quality and dynamics of client-therapist interaction, and it can serve as a key input for systems mediating such interactions. Motivational interviews exhibit three important characteristics. First, there are two distinct roles, namely client and therapist. Second, they are often highly emotionally charged, which can be expressed both in text and in prosody. Finally, context is of central importance to classify any given utterance. Previous works did not adequately incorporate all of these characteristics into utterance classification approaches for mental health dialogues. In contrast, we present M^{3}TCM, a Multi-modal, Multi-task Context Model for utterance classification. Our approach for the first time employs multi-task learning to effectively model both joint and individual components of therapist and client behaviour. Furthermore, M^{3}TCM integrates information from the text and speech modality as well as the conversation context. With our novel approach, we outperform the state of the art for utterance classification on the recently introduced AnnoMI dataset with a relative improvement of 20% for the client- and by 15% for therapist utterance classification. In extensive ablation studies, we quantify the improvement resulting from each contribution.

Keywords: utterance classification, multi-task learning, conversation context, multi-modal, motivational interviewing

1. Introduction

Motivational interviewing (MI) is an important tool in helping clients to achieve goals such as reducing alcohol consumption and smoking, managing asthma or diabetes, or increasing physical activity. Automatic analysis of motivational interviewing has on the one hand the potential to improve our understanding of the effectiveness of different techniques. On the other hand, it is also a basis for building social agents that can meaningfully interact with clients. To this end, automatic approaches need to be able to precisely categorize the utterances of both counselor and client.

Motivational interviews have three important characteristics. First, client and therapist have distinct roles, including different sets of ground truth utterance labels (Wu et al., 2022a). Second, motivational interviews are often emotionally charged. Third, conversation context is crucial to interpret any given utterance. Previous approaches to utterance classification in motivational interviews did not fully take advantage of all of these characteristics. While approaches integrating text and audio do exist (Aswamenakul et al., 2018; Gupta et al., 2014; Singla et al., 2018; Tavabi et al., 2020), they commonly do not model the conversation context. The few approaches that do model conversation context, are either not multi-modal (Tavabi et al., 2020), or only considered a single utterance as context (Gupta et al., 2014). Most importantly, all previous approaches addressed patient and therapist utterance classification in completely separate models. The potential benefit of multi-task learning remains unexplored.

To overcome these limitations, we present M^{3}TCM, a multi-modal multi-task context model for utterance classification in motivational interviewing. M^{3}TCM for the first time uses multi-task learning to effectively model both joint and individual components of the two tasks of classifying therapists' and clients' utterances. Our approach furthermore effectively leverages prosodic information as well as the conversation context. In evaluations on the recently introduced AnnoMI dataset (Wu et al., 2022a), M^{3}TCM outperforms previously proposed approaches by a significant margin (0.66 F1 vs. 0.55 F1 for client utterances, 0.83 vs. 0.72 F1 for therapist utterances). We present extensive ablation experiments, documenting the importance of the multi-task framework and of utilizing text and audio modalities in conjunction with conversation context. We furthermore for the first time evaluate different sizes of the input window, showing that the optimal context size is significantly larger than those used in previous work.

2. Related Work

Our work is related to utterance classification in mental health conversations and to multi-task learning approaches applied to conversation analysis.

LREC-COLING 2024, pages 10872–10879
20-25 May, 2024.

2.1. Utterance Classification in Mental Health Conversations

Ewbank et al. (2020) classified therapist utterances obtained from transcripts of Cognitive Behaviour Therapy (Brewin, 2006) sessions into 24 categories to predict therapy outcome. In another study, Ewbank et al. (2021) employed deep learning techniques to automatically classify patient talk types within Cognitive Behaviour Therapy.

Previous approaches confirm the importance of fusing text and audio information for utterance classification in MI (Aswamenakul et al., 2018; Tavabi et al., 2020; Gupta et al., 2014; Singla et al., 2018). Most approaches only address the problem of client talk type classification, but Singla et al. (2018) proposed an approach based on single utterances that is applied to therapist and client talk type classification, integrating text and audio information. Only a subset of utterance classification approaches modeled the conversation context in order to classify a target utterance. Tavabi et al. (2020) took 3 previous text-utterances from both client and therapist as context to classify current client utterance. Gupta et al. (2014) investigated the effect of laughter and prosodic differences in MI interviews, using the previous therapist's utterance as context. In summary, while several approaches integrated text and audio modalities, these commonly do not explore the effect of the size of the context. The few approaches that do model conversation context do not provide analyses on the impact of the size of the input window. Crucially, none of the existing approaches leverages a multi-task learning framework to simultaneously learn models for therapist and client utterances.

Recently Wu et al. (2022a) introduced AnnoMI, an expert-annotated dataset of motivational interviews available on Youtube. The dataset derives its annotations from the Motivational Interviewing Skills Code (MISC) (Miller and Rollnick, 2012) and has a different set of labels for client and therapist. We use AnnoMI because it is the biggest publicly available dataset with MI interviews, annotated by experts. Existing work on this dataset employed language models to create separate, single-utterance text-based classifiers for therapist and client utterances (Wu et al., 2022b, 2023a). To the best of our knowledge, we present the first multi-modal, context-aware, multi-task approach to utterance classification on the AnnoMI corpus.

2.2. Multi-Task Approaches

Previous work applied multi-task learning for dialogue analysis in several setups. Ide and Kawahara (2021) proposed a multi-task learning method for emotion-aware dialogue response generation, emphasizing the synergy between generation and classification tasks. They train the same model to generate dialogue responses and at the same time detect emotion. Liu et al. (2022) introduced EmoDM, which at the same time learns to track emotional states and empathetic dialogue policy selection. Kollias (2022) presented the ABAW Competition, which includes challenges like Valence-Arousal Estimation and Expression Classification using multi-task learning on the Aff-Wild2 database. In a subsequent iteration, Kollias (2022) highlighted the potential of multi-task approaches in emotion detection and classification using both synthetic data and multi-task learning to classify valance or arousal and emotions. To the best of our knowledge, multi-task learning was not yet applied to model the different roles speakers have in motivational interviewing.

3. Method

Figure 1 illustrates the architecture of our model. Text- and audio embeddings are extracted from k consecutive utterances of therapist and client. A shared self-attention layer is used to model conversation context across utterances, and task-specific classification networks are utilized to produce classification outputs for the therapist and client.

3.1. Input Embeddings

In the following we discuss how we obtained per-utterance embeddings from text and audio inputs. For text data, we used RoBERTa Large (Liu et al., 2020). RoBERTa, short for "Robustly optimized BERT approach," is a variant of the BERT model designed for natural language processing. RoBERTa improved BERT's performance by altering the training regimen, notably removing the next-sentence prediction objective and utilizing dynamic masking for more efficient pre-training. The model was trained with more data and larger batch sizes, resulting in improved accuracy and demonstrating the significance of meticulous training details. RoBERTa achieved state-of-the-art results in various NLP benchmarks (Liu et al., 2020) including emotion (Adoma et al., 2020) and depression (Gupta et al., 2023) recognition.

To encode prosodic information, we made use of the Audio Spectrogram Transformer (AST) (Gong et al., 2021), a specialized model designed to handle audio classification tasks using the transformer architecture. AST directly operates on audio spectrograms and achieved state-of-the-art results on recognizing human speech (Gemmeke et al., 2017a), command (Warden, 2018) and also difference between human and environmental sounds (Piczak, 2015). The Audio Spectrogram Transformer (AST) (Gong et al., 2021) is particularly well-suited for analysing prosody due to its ability to model inter-dependencies across time and thereby extract intricate patterns from audio data. AST's

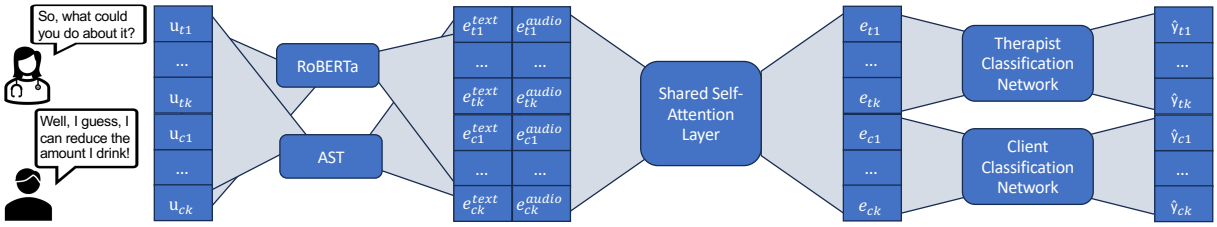

Figure 1: Overview over the M^{3}TCM Model. Several consecutive therapist and client utterances (u_{ti} and u_{ci}, respectively) are encoded using RoBERTa and AST models, producing text and audio embeddings. A shared self-attention layer models conversation context across utterances. Finally, separate classification networks produce predictions for therapist and client utterances.

attention-based mechanism allows it to focus on specific aspects of the audio spectrogram, such as the variations in pitch, tempo, and volume, which are integral components of prosody.

3.2. M^{3}TCM Approach

Our model processes k utterances of each patient and therapist in parallel. At each time step $i = 1..k$ we have the textual utterances of the therapist and client, denoted as u_{ti} and u_{ci} respectively, along with their corresponding audio spectrograms. RoBERTa is used to extract $2k$ text embeddings $E_{text} = e^{text}_{t1}..e^{text}_{tk}, e^{text}_{c1}..e^{text}_{ck}$ from therapist and client utterances. AST on the other hand produces the corresponding audio embeddings $E_{audio} = e^{audio}_{t1}..e^{audio}_{tk}, e^{audio}_{c1}..e^{audio}_{ck}$.
Text and audio embeddings for any given utterance are subsequently concatenated:

$$E = E_{audio} \oplus E_{text} \quad (1)$$

which results in the combined embeddings $E = e_{t1}..e_{tk}, e_{c1}..e_{ck}$. To incorporate conversation context in our classification approach, we model relations between utterances with a self-attention layer (Vaswani et al., 2017):

$$E' = SelfAttention(E) \quad (2)$$

Our multi-task learning approach employs two task-specific networks working on separate subsets of the $2k$ transformed embeddings $E' = e'_{t1}..e'_{tk}, e'_{c1}..e'_{ck}$. The therapist utterance classification network f_t receives as input the first k embeddings $E'_{therapist} = e'_{t1}..e'_{tk}$ and outputs k classification decisions $\hat{y}_t = \hat{y}_{t1}..\hat{y}_{tk}$, one for each therapist input utterance:

$$\hat{y}_t = f_t(E'_{therapist}) \quad (3)$$

Analogously, the client classification network f_c produces predictions $\hat{y}_c = \hat{y}_{c1}..\hat{y}_{ck}$ for the k client input utterances. While client and therapist classification networks have separate weights, the self-attention layer is shared between both tasks. This allows our self-attention layer to learn both task-dependent and task-independent aspects of behaviour. To be precise, the multi-task-learning takes place through query-key interactions across client and therapist utterances.

3.3. Implementation Details

To address the class imbalances(Sub-section 4.1) resulting from client and therapist behaviour, we use the Focal Loss function, suitable for imbalanced classification scenarios Lin et al. (2017). M^{3}TCM Shared layer has the dimension of 1551×1024. Both client and therapist specific heads have two layers, with 1024×512 and 512×256 dimensions. To improve reproducibility, we make our code publicly available[1].

4. Experiments

4.1. Data Prepossessing

We began with the AnnoMI dataset from Wu et al. (2023a), consisting of 13551 utterances transcribed from 133 Youtube videos. Since the initial publishing of AnnoMI, some of those videos have been removed from Youtube, our dataset contained in the end 125 videos. Given our multi-modal approach including audio, we had to remove utterances of non-available videos, leaving us with 12778 instances, 6338 for client and 6440 for therapist. To extract the per-utterance audios, we isolated audio from videos and segmented them using the utterance timestamps provided with AnnoMI. Instances with multiple annotators were harmonized by selecting the most frequent annotation.
Our targets were the client talk type class and the main therapist behaviour. The client class was imbalanced: 63% “neutral”, 25% “change”, and 12% “sustain”. On the other hand, the therapist’s class distribution showcased a more even spread: 31% for “other”, 29% for “question”, 25% for “reflection”, and 15% for “therapist_input”.

[1] https://git.opendfki.de/philipp.mueller/m3tcm

Models	Client				Therapist				
	Average	Change	Neutral	Sustain	Average	Reflection	Question	Input	Other
Random Baseline	0.33	0.25	0.63	0.12	0.25	0.25	0.29	0.15	0.31
Wu et al. (2023a)	0.55	0.51	**0.74**	0.39	0.72	0.77	0.86	0.63	0.64
M^{3}TCM Without Finetuning	0.54	0.70	0.42	0.41	0.73	0.65	0.82	0.81	0.63
M^{3}TCM Text Only Single Task	0.58	0.76	0.56	0.43	0.77	0.73	0.86	0.82	0.68
M^{3}TCM Audio Only Single Task	0.40	0.65	0.38	0.18	0.44	0.40	0.60	0.44	0.31
M^{3}TCM Audio Only No Context	0.38	0.65	0.36	0.13	0.40	0.38	0.58	0.40	0.25
M^{3}TCM Text Only No Context	0.57	0.73	0.52	0.45	0.77	0.74	0.86	0.82	0.67
M^{3}TCM Audio Only	0.46	0.73	0.43	0.21	0.49	0.46	0.68	0.48	0.33
M^{3}TCM Text Only	0.63	0.80	0.59	0.49	0.80	0.76	0.89	0.85	0.68
M^{3}TCM No Context	0.61	0.78	0.57	0.48	0.76	0.70	0.83	0.87	0.65
M^{3}TCM Single Task	0.60	0.78	0.57	0.46	0.77	0.70	0.85	0.87	0.65
M^{3}TCM	**0.66**	**0.83**	0.62	**0.52**	**0.83**	**0.81**	**0.89**	**0.88**	**0.73**

Table 1: Classification results for M^{3}TCM compared to baselines and ablation conditions. We report per-class, as well as macro-averaged F1 scores for both client and therapist classification tasks.

4.2. Training Details

We used 5 Fold Cross Validation stratified by video to guarantee that no utterances form the same video can appear both in train and test sets. We used $\frac{3}{5}$ of the data for training, $\frac{1}{5}$ for validation, and $\frac{1}{5}$ for testing.

In a first step, we fine tuned both AST (Gong et al., 2021) and RoBERTA Large (Liu et al., 2020) model on our dataset. We also tried using AST and RoBERTa without finetuning, but that led to inferior results. In a second step, we trained the full M^{3}TCM model for 100 epochs and choose the best model based on the performance on the validation set. One thing to note is that at this stage of the training the weights of the finetuned RoBERTa and AST layer were fixed and as we said before was selected based on the best performance on the validation set.

Both for the finetuning and the final training phase we used the the AdamW optimizer at a learning rate of 1e-5 Loshchilov and Hutter (2019) and trained for 100 epochs. We selected the best model from these 100 epochs by evaluating F1 score on the validation set.

5. Results

In line with previous work (Wu et al., 2022b, 2023a), we evaluated all approaches using the F1 score. We do so both with per-class F1 scores as well as separate macro-averaged F1 scores for the patient and therapist utterance classification tasks. In Table 1, we report results for M^{3}TCM as well as baselines and ablation conditions.

M^{3}TCM outperforms all other approaches, reaching 0.66 F1 for the client and 0.83 F1 for therapist utterance classification. This is a clear improvement over the previous state of the art by Wu et al. (2023a) (0.55 F1 client, 0.72 F1 therapist). Crucially, our ablation experiments confirm the utility of multi-task learning. Models trained separately on patient and therapist utterance classification ("M^{3}TCM Single Task"), only reached 0.60 F1 for client and 0.77 F1 for therapist. Furthermore, consistent improvements over mono-modal ablations ("M^{3}TCM Audio Only / Text Only") document the utility of fusing text and audio information. In addition, we observed that the inclusion of conversation context leads to clear improvements: we see that without context it achieves for client F1 of 0.61 and for therapist 0.76.

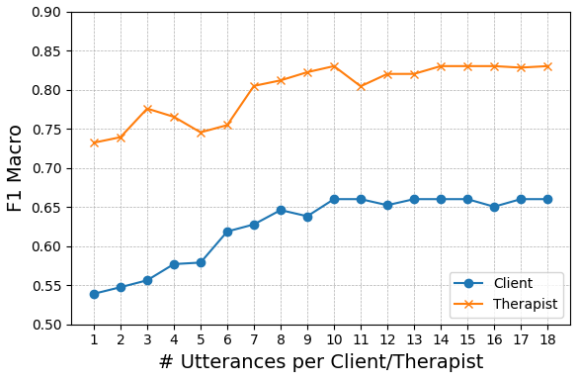

Figure 2: Performance for therapist and client utterance classification for different context sizes.

Our M^{3}TCM model has a slightly lower F1 score (0.62) for the majority "neutral" class for client talk type compared to random guessing (0.63 F1). The reason for this is that we decided to optimise our model to perform well on all classes (and not primarily on the majority class), which is reflected in consistently higher scores for the minority classes. For "change", M^{3}TCM reached 0.83 F1 versus 0.51 F1 for Wu et al. (2022b), and 0.25 F1 for the random baseline. For the challenging minority class "sustain", M^{3}TCM reached 0.52 F1 versus 0.39 F1 for Wu et al. (2022b), and 0.12 F1 for the random

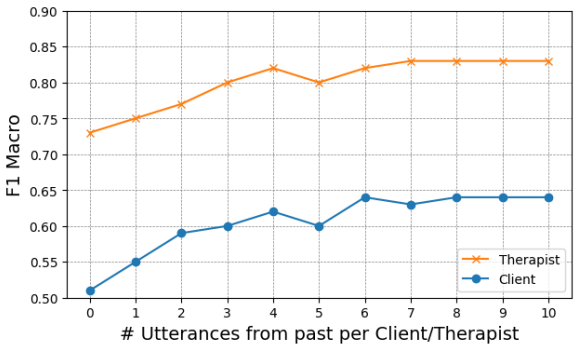

Figure 3: Performance of therapist and client utterance classification for different context sizes in an online evaluation scenario.

baseline. A precise distinction between "change" and "sustain" is especially important in motivational interviews, as these are highly informative classes concerning behaviour change.

To better understand the utility of conversation context, we conducted an experiment with varying numbers of utterances as context (see Figure 2). We observed a clear increase in F1 score for both therapist and patient when increasing the number of patient/therapist utterances we input to the model from 1 to 10. For more than 10 utterances performance reaches a plateau, while memory utilization continues to increase. We therefore determine 10 utterances per patient/therapist as the optimal input size, which is much larger than the input window of maximally 3 utterances used in previous work (Tavabi et al., 2020).

It is important to note that our model is evaluated in an offline scenario, i.e. for the classification of a given utterance it also has access to future utterances. To understand its capabilities in an online classification setup, we analyze the prediction performance when only using the prediction on the last utterance of the input window. We present the corresponding results for varying sizes of the input window (i.e. previous) utterances in Figure 3. In general, the performance is very similar to the offline approach, demonstrating the utility of our approach in online classification scenarios.

6. Conclusion and Future Work

In this work, we presented M^{3}TCM, a multi-modal and context-sensitive approach to utterance classification in motivational interviews that for the first time leverages multi-task learning to model both therapist and patient at the same time. We showed clear improvements over the previous state of the art as well as ablated versions of our model. As such, our work underlines the importance of models that make use of all the available information to build highly accurate conversation analysis systems. For future work, it would be interesting to integrate the video modality alongside text and prosody. Furthermore, our multi-task approach could be applied to different scenarios that exhibit asymmetrical roles in conversation. These may include psychiatric interactions (König et al., 2022), sales conversations, teacher-student interactions (Cafaro et al., 2017), or police interrogations. In addition, it will be interesting to integrate predicted utterance classes as input features in nonverbal conversational behaviour generation approaches (Withanage Don et al., 2023).

7. Acknowledgements

J. Alexandersson and P. Müller were funded by the European Union Horizon Europe programme, grant number 101078950.

8. Bibliographical References

Acheampong Francisca Adoma, Nunoo-Mensah Henry, and Wenyu Chen. 2020. Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. In *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 117–121.

Chanuwas Aswamenakul, Lixing Liu, Kate B Carey, Joshua Woolley, Stefan Scherer, and Brian Borsari. 2018. Multimodal analysis of client behavioral change coding in motivational interviewing. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 356–360.

David C Atkins, Mark Steyvers, Zac E Imel, and Padhraic Smyth. 2014. Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. *Implementation Science*, 9(1):1–11.

John S Baer, Elizabeth A Wells, David B Rosengren, Bryan Hartzler, Blair Beadnell, and Chris Dunn. 2009. Agency context and tailored training in technology transfer: A pilot evaluation of motivational interviewing training for community counselors. *Journal of substance abuse treatment*, 37(2):191–202.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66.

Brian Borsari, John TP Hustad, Nadine R Mastroleo, Tracy O'Leary Tevyaw, Nancy P Barnett, Christopher W Kahler, Erica Eaton Short, and Peter M Monti. 2012. Addressing alcohol use and problems in mandated college students: a randomized clinical trial using stepped care. *Journal of consulting and clinical psychology*, 80(6):1062.

Chris R Brewin. 2006. Understanding cognitive behaviour therapy: A retrieval competition account. *Behaviour research and therapy*, 44(6):765–784.

Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel Valstar. 2017. The noxi database: multimodal recordings of mediated novice-expert interactions. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 350–359.

Kate B Carey, James M Henson, Michael P Carey, and Stephen A Maisto. 2009. Computer versus in-person intervention for students violating campus alcohol policy. *Journal of consulting and clinical psychology*, 77(1):74.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Michael P Ewbank, Ronan Cummins, Valentin Tablan, Sarah Bateup, Ana Catarino, Alan J Martin, and Andrew D Blackwell. 2020. Quantifying the association between psychotherapy content and clinical outcomes using deep learning. *JAMA psychiatry*, 77(1):35–43.

MP Ewbank, R Cummins, V Tablan, A Catarino, S Buchholz, and AD Blackwell. 2021. Understanding the relationship between patient language and outcomes in internet-enabled cognitive behavioural therapy: A deep learning approach to automatic coding of session transcripts. *Psychotherapy Research*, 31(3):300–312.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017a. Audio set: An ontology and human-labeled dataset for audio events. pages 776–780.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017b. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780.

Yuan Gong, Yu-An Chung, and James Glass. 2021. AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*, pages 571–575.

Khushi Gupta, Razaq Jinad, and Qingzhong Liu. 2023. Comparative analysis of nlp models for detecting depression on twitter. In *2023 International Conference on Communications, Computing and Artificial Intelligence (CCCAI)*, pages 23–28.

Rahul Gupta, Panayiotis G Georgiou, David C Atkins, and Shrikanth S Narayanan. 2014. Predicting client's inclination towards target behavior change in motivational interviewing and investigating the role of laughter. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

C Howes, M Purver, and R McCabe. 2013. Using conversation topics for predicting therapy outcomes in schizophrenia. *Biomed Inform Insights*, 6(Suppl 1):39–50.

Tatsuya Ide and Daisuke Kawahara. 2021. Multi-task learning of generation and classification for emotion-aware dialogue response generation. In *Proceedings of the NAACL Student Research Workshop*.

Mélanie Jouaiti and K. Dautenhahn. 2022. Dysfluency classification in stuttered speech using deep learning for real-time applications. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing*.

D. Kollias. 2022. Abaw: Learning from synthetic data & multi-task learning challenges. In *European Conference on Computer Vision Workshops*, pages 157–172.

Alexandra König, Philipp Müller, Johannes Tröger, Hali Lindsay, Jan Alexandersson, Jonas Hinze, Matthias Riemenschneider, Danilo Postin, Eric Ettore, Amandine Lecomte, et al. 2022. Multimodal phenotyping of psychiatric disorders from social interaction: Protocol of a clinical multicenter prospective study. *Personalized Medicine in Psychiatry*, 33:100094.

Vivek Kumar, Simone Balloccu, Zixiu Wu, Ehud Reiter, Rim Helaoui, Diego Recupero, and Daniele Riboni. 2023. Data augmentation for reliability and fairness in counselling quality classification. In *1st Workshop on Scarce Data in Artificial Intelligence for Healthcare-SDAIH, INSTICC; SciTePress: Setúbal, Portugal*, pages 23–28.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 2980–2988.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Roberta: A robustly optimized bert pretraining approach.

Yuhan Liu, Jun Gao, Jiachen Du, Lanjun Zhou, and Ruifeng Xu. 2022. Emodm: Empathetic response generation with emotion-aware dialogue management.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations*.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

William R Miller and Stephen Rollnick. 2012. *Motivational interviewing: Helping people change*. Guilford press.

Y. Nakano, E. Hirose, T. Sakato, S. Okada, and Jean-Claude Martin. 2022. Detecting change talk in motivational interviewing using verbal and facial information. In *Proceedings of the ACM International Conference on Multimodal Interaction*, pages 5–14.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2016. Building a motivational interviewing dataset. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 42–51.

Karol J Piczak. 2015. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

K Singla, Z Chen, N Flemotomos, J Gibson, D Can, DC Atkins, and S Narayanan. 2018. Using prosodic and lexical information for learning utterance-level behaviors in psychotherapy. *Interspeech*, 2018:3413–3417.

Leili Tavabi, Kalin Stefanov, Larry Zhang, Brian Borsari, Joshua D Woolley, Stefan Scherer, and Mohammad Soleymani. 2020. Multimodal automatic coding of client behavior in motivational interviewing. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 406–413.

Lucas Teoh, A. Ihalage, Srooley Harp, Zahra F. Al-Khateeb, A. Michael-Titus, J. Tremoleda, and Yang Hao. 2022. Deep learning for behaviour classification in a preclinical brain injury model. *PLOS ONE*, 17(4).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Pete Warden. 2018. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*.

Daksitha Senel Withanage Don, Philipp Müller, Fabrizio Nunnari, Elisabeth André, and Patrick Gebhard. 2023. Renelib: Real-time neural listening behavior generation for socially interactive agents. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 507–516.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2023a. Creation, analysis and evaluation of annomi, a dataset of expert-annotated counselling dialogues. *Future Internet*, 15(3):110.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. 2022a. Anno-mi: A dataset of expert-annotated counselling dialogues. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6177–6181.

Zixiu Wu, Simone Balloccu, Ehud Reiter, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2023b. Are experts needed? on human evaluation of counselling reflection generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6906–6930.

Zixiu Wu, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2022b. Towards automated counselling decision-making: Remarks on therapist action forecasting on the annomi dataset. *Change*, 25:17.

Zhihan Zhang, Wenhao Yu, Mengxia Yu, Zhichun Guo, and Meng Jiang. 2022. A survey of multi-task learning in natural language processing: Regarding task relatedness and training methods. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*.